MS/MS Spectrum Prediction for MHC-Associated Peptides with a Fine-Tuned Model

by

Zhenbo Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

To improve the quality of spectral library search, several MS/MS spectrum predictors have been developed in the last decades. After success in various fields, deep learning techniques are adopted by MS/MS spectrum predictors to increase the accuracy of predicted spectra. However, the quality and quantity of the training set are both required to train a deep learning model. Due to the less representation of MHC-associated peptides in most spectral libraries, current MS/MS spectrum predictors provide less accurate predicted spectra for MHC-associated peptides than their performance for other peptides.

In this thesis, we built several MHC-associated peptide spectral libraries for training and evaluation purposes. We selected PredFull as our base model and performed transfer learning with these MHC-associated peptide libraries, which are much smaller than common tryptic spectral libraries. The result showed that the fine-tuned model outperformed the original model significantly when predicting MHC-associated peptides.

Acknowledgements

I am deeply grateful to extend my sincerest appreciation to Professor Bin Ma and Professor Yang Lu for their unwavering guidance, invaluable assistance, and relentless encouragement throughout the entirety of this project. Their expertise and mentorship have been invaluable in helping me throughout this journey. Furthermore, I would like to express my heartfelt gratitude to my spouse, Ran Ye, whose unwavering love accomplished me throughout my graduate journey. Their presence, especially during the challenging moments, has given me the strength and motivation to overcome obstacles. I am profoundly grateful for their persistent belief in me and endless support. I would also like to thank our dog Hakobi, who has been reminding me to do outdoor exercises even if I have a very tight schedule.

Dedication

This is dedicated to my spouse, Ran Ye.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

Dedication v

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Background . . . 1

1.2 Workflow of LC-MS/MS . . . 2

1.3 Peptide Identification From MS/MS Spectra . . . 2

1.4 Traditional Prediction Tools for MS/MS spectra . . . 3

1.5 Deep Learning and its Application on MS/MS Spectrum Prediction . . . 4

1.6 MHC Associated Peptides . . . 4

1.7 Transfer Learning . . . 5

1.8 Contribution . . . 5

2 Transfer Learning Methods 7

2.1 Selection of the Base Model 7

2.2 Representation of MS/MS Spectrum 8

2.3 Fine-tuning and Hyper-parameters 9

2.3.1 Trainable Layers 9

2.3.2 Optimizer 9

2.3.3 Loss Function 9

2.3.4 Batch size 10

2.3.5 Epochs 10

2.3.6 Learning Rate 10

2.4 Score Function 10

2.5 HardWare and Software Environment 11

3 Data Source and Data Process 12

3.1 MS Data Selection 12

3.1.1 MHC-I Q-Exactive Model 12

3.1.2 MHC-I Fusion Lumos and Eclipse Tribrid Model 13

3.1.3 MHC-II Model 13

3.2 File format conversion 13

3.3 Database Search 14

3.4 Dataset Split 14

3.4.1 Q-Exactive Family 15

3.4.2 Fusion Lumos and Eclipse Tribrid 15

3.4.3 MHC II-Associated Peptides 16

4 Result 17

4.1 Q-Exactive Model 17

4.1.1 Performance on Q-Exactive family datasets with 5-fold Cross validation 18

4.1.2 Performance on New Q-Exactive Family Datasets 18
4.1.3 Pairwise Performance Evaluation on Q-Exactive Family Datasets . 19
4.1.4 Case Study of Predicted MS/MS Spectrum Before and After Fine-Tuning 19
4.2 Prediction Performance on Fusion Lumos and Eclipse Tribrid Datasets . . 23
4.3 Prediction Performance on MHC II-Associated Peptides 24

5 Factor of Improvements Analysis 26
5.1 Tryptic peptides were more likely to be improved by fine-tuning 26
5.1.1 The original model's performance dropped with non-tryptic models 27
5.1.2 Fine-Tuning Improved Performance on both Tryptic and Non-Tryptic Peptides 27
5.2 Effect of amino acids with possitive charges(K,R,H) 28
5.3 Effect of Peptide Charges 29
5.4 Effect of Proline 29
5.5 Number of Annotated peaks in the Experimental Spectrum 30
5.5.1 Definition of Less Peak Peptides 30
5.5.2 Performance with Less Peak Peptides 31
5.6 Summary 31

6 Conclusion 34

References 36

List of Figures

2.1 Architecture of the neural network. PredFull as the backbone model. . . . 7

4.1 Five-fold cross-validation on PXD004894. The dotted lines represent the quartile (25%,50%,75%) of the distribution. V1: partitions P2,P3,P4,P5 were used as the training set, and partition P1 were used as the validation set. 17

4.2 The dotted lines represent the quartile (25%,50%,75%) of the distribution. The violin plot showed the performance on new datasets tested on the V2 model. 18

4.3 Pairwise comparison of each observation in the four new datasets. S^2 was the square of cosine similarity between the real spectrum and the predicted spectrum. For each observation, X-axis represented S^2 for the original model before fine-tuning, and Y-axis represented the fine-tuned model. 20

4.4 Example of predictions improved after fine-tuning. In each subfigure, the upper part was the predicted spectrum and lower part was the real mass spectrum. 21

4.5 Example of predictions with no improvements after fine-tuning. In each subfigure, the upper part was the predicted spectrum and lower part was the real mass spectrum. 22

4.6 Five-fold cross-validation on Fusion Lumos and Eclipse Tribrid datasets. The dotted lines represent the quartile (25%,50%,75%) of the distribution. V1 means we used partitions P2,P3,P4,P5 as the training set, and partition P1 as the validation set. 23

4.7 Evaluation on Independent Test Set of Fusion Lumos and Eclipse Tribrid . 24

4.8 Three-fold cross-validation on MHC II-associated peptides from PXD004894. The dotted lines represent the quartile (25%,50%,75%) of the distribution. V1 means we used partitions P2,P3 as the training set, and partition P1 as the validation set. 25

4.9 Evaluation on Independent Test Set of MAP II-associated peptides 25

5.1 Distribution of prediction score in the Independent Test Set. Score function was the cross similarity of the predicted spectrum and the experimental(real) spectrum 28

5.2 The relationship between the number of Lys, Arg and His in the peptide and the models' prediction score 29

5.3 Prediction Score for different peptides' charges. Left: Without Fine-Tune. Right: With Fine-Tune 30

5.4 Effect of Proline. 31

5.5 Low-peaks means the real spectrum had less than 25 annotated peaks . . . 33

List of Tables

3.1 The number of PSMs in each MHC-I peptide datasets 13
3.2 Grouping PXD004894 into 5 Partitions and an Independent Test Set by Patients 15
3.3 The number of PSMs and Bags of Fusion Lumos and Eclipse Tribrid datasets 16

Chapter 1

Introduction

1.1 Background

Proteins are the fundamental macromolecule for all living molecules, involving in the constitution of organs and muscles, synthesis of nuclei acids, hormones and antibodies that defence invading pathogens, etc. In 1995, Marc Wilkins created the term ”proteomics” [1] , which means studying the overall protein content of a cell.[2] Proteomics is a cross-disciplinary research field and in turn plays an crucial role in various areas, including studying protein-protein interactions, analyzing post-translational modifications (PTMs) [3], identifying biomarkers for various diseases like cancer and drug discovery. [1]

Identification of proteins and peptides is considered as the pivotal aspect of proteomics. The state-of-the-art technique of this process is known as tandem mass spectrometry (MS/MS). In general, proteins or peptides are separated and measured into the mass-to-charge ratio of ions to obtain identification and quantification. [4] Next, the acquired MS/MS spectra are matched with sequence in protein or peptide database based on specific algorithms. The success of spectral library searching depends on the existence of a high-quality spectral library covering the spectra of the peptides to be identified.

To improve the quality of spectral library search, several MS/MS spectrum predictors for accepting peptide sequences as input, and generating predicted MS/MS spectra as output have been developed, in which deep learning techniques greatly increase the accuracy of predicted spectra. However, for the study of MHC(Major histocompatibility complex)-associated peptides, which are essential for immune system, there is a lack of high-quality presentation for most spectral libraries. In particular, current MS/MS spectrum predictors provide less accurate predicted spectra for MHC-associated peptides than their performance for other peptides.

1.2 Workflow of LC-MS/MS

The typical MS/MS workflow for protein identification begins at sample preparation.[4] Traditionally, the mixture of protein would be separated by their physical or chemical properties at the beginning to reduce the further analysis complexity, named "top-down". [3] To the contrary, another method called "bottom-up", is to apply a protease(usually trypsin) that breaks the peptide bonds of protein before separating simple or complex protein samples with gel electrophoresis or liquid chromatography. When "bottom-up" method is applied on the complex protein mixture, it is also called "Shotgun"[4].

After digestion with protease, there will be a variety of peptides needing to be separated before MS/MS. Liquid chromatography (LC) is the most widely used separation method to feed these peptides into MS/MS based on the different retention time(RT) of each peptide. After that, samples will go through MS/MS.

Firstly, these peptides will be ionized by ion source to obtain peptide's individual mass-over-charge ratio. The first MS analyzer could recognize these precursor ions(parent ions) [5], separate them by mass-over-charge(m/z) ratio, and output MS1 spectra. Meanwhile, the selected precursor ions will undergo fragmentation and transform into product ions. The second MS analyzer could recognize them and output MS2 spectra.[6]

1.3 Peptide Identification From MS/MS Spectra

In this section, we introduced several methods for peptide identification from MS/MS spectra. In 1994, Eng et al. published a software SEQUEST to identify peptide sequences from MS/MS spectra if they existed in a protein database. [7] For each MS/MS spectrum, the algorithm would find out all the peptides with similar mass as candidates from the protein database. Considering the properties of the instruments, a mass tolerance would be specified. Then, a theoretical MS/MS spectrum would be generated for each peptide candidate, and a scoring function would be applied on the theoretical MS/MS spectrum and the experimental MS/MS spectrum. [7]. This method, called database search later, became the standard method for identifying peptides[8]. After this, various database search engines have been published. For example, Comet, released in 2013, facilitated a fast cross-correlation algorithm as a filter to reduce the computation amount. The algorithms behind it were implemented with multi-thread support that increaseing its efficiency on modern multi-core CPUs[9]

TANDEM, published by Craig et al. in 2004, searched peptide candidates from spectral libraries that included both peptide sequences and their annotated MS/MS spectra, instead of protein databases only containing protein sequences.[10] When performing a spectral library search, the experimental spectrum would be compared with annotated spectra from the library, generating scores representing the similarity between them[11]. The scoring function, based on equations proposed by Fenyö et al., considered both the number of matched peaks and the intensities of these peaks[12]. In 2007, Lam et al. published a new spectral library search engine called SpectraST. They designed a new set of equations to calculate the score between experimental spectrum and the spectra from the library. Compared to the widely-used database search engine SEQUEST, SpectraST reached a better sensitivity at the same FDR level. [13]. It was integrated into Trans Proteomic Pipeline (TPP) suite[14], and soon became the most popular spectral library search engine.

1.4 Traditional Prediction Tools for MS/MS spectra

The success of spectral library searching depends on the existence of a high-quality spectral library, which covers the spectra of the peptides to be identified. [15] A high-accuracy MS/MS spectra predictor could improve the accuracy of peptide identification. [16] In this section, we introduced various predictors released in the past two decades.

In 2004, MassAnalyzer showed its advantages in the low-energy collision-induced dissociation (CID) spectra prediction based on a kinetic model, consisting of 11 potential fragmentation pathways including ”cleverages of backbone amide bonds” that could result in b and y ions[17]. There were 236 parameters during fragmentation process, which were optimized by facilitation of a CID dataset with 5605 spectra. The program was implemented in C++, reaching a considerable speed for spectrum prediction ”on a 1.7 GHz Pentium IV computer”.

In 2006, Arnold et al. published PeptideArt, a model with ”two-layer feed-forward neural networks”[18], outperforming the kinetic model MassAnalyzer[19]. MS2PIP, released by Sven Degroeve et al. in 2013, used random forests regression to predict the peak intensity that showed better accuracy than PeptideArt[19].

1.5 Deep Learning and its Application on MS/MS Spectrum Prediction

Deep learning (DL) had became the most widespread machine learning (ML) application since last decade. Compared to traditional methods, deep learning abandoned human-designed rules, but built a multi-layer artificial neural network (ANN). Data were fed into this network to adjust its weights[20] . Several researches showed that building a deep learning model requires a larger dataset to reach expected performance [21].

Inspired by genomics, a public repository for proteomics data, ProteomeXchange, went live in 2006[22] [23]. It collected the MS/MS spectrum data uploaded by researchers from all over the world, and these datasets became public available for further research. It eliminated hassles for researchers searching for proteomics data, especially raw MS/MS spectra, and benefited training deep learning models with these datasets.

Since then, more and more high quality MS/MS spectrum datasets were published. For example, NIST Consensus Human HCD Libraries was initially published in 2016 and underwent several updates[24]. After the last update in 2020, 600k unique peptide sequences and their associated spectra were collected. NIST Human Synthetic HCD library , published in 2017, contained 189k unique peptides [25]. All of these pubic spectral libraries achieved high quality of MS/MS spectrum and extensive coverage of peptides.

An ideal process of training deep learning models was considered being easy, concise and quick. Machine learning frameworks like Tensorflow[26] and PyTorch[27] provided a user-friendly interface and were successfully applied in various fields, including prediction of MS/MS spectra. For example, In 2019, Guan el at. applied bidirectional LSTM for spectra prediction[28] and Siegfried Gessulat et al. Published Prosit [29] that was composed of a bi-directional recurrent neural network, a recurrerrent GRU layer and an attention layer.

1.6 MHC Associated Peptides

Major histocompatibility complex (MHC; also known as human leukocyte antigen(HLA) in humans) molecules play an important role in immune response [30][31]. MHC molecules could display a small peptide fragment derived from pathogens or self-proteins, which is also called MHC-associated Peptides(MAP), on the surface of a cell for T cells recognition. [32] Generally,MHC molecules can be grouped into MHC Class I, MHC Class II and MHC Class III, and the first two classes attracted most of the attention. [33] MHC Class I-associated peptides and MHC Class II-associated peptides differ in both peptide structure and their role in the intracellular pathways. Most MHC Class I peptides have length between 9-12 amino acids, which came from the ubiquitin-proteasome system degradating intracellular proteins. Cytotoxic CD8+ T cells recognizes MHC Class I-peptides complex to enable immune system against pathogens and cancer.[34]. While For MHC class II, associated peptides are typically 13-17 amino acid long and are recognized by Cytotoxic CD4+ T cells[32] .

1.7 Transfer Learning

The goal of transfer learning is to learn features from the source domain data and facilitate our learning to improve the prediction on the target domain data. Typically, the size of target domain was much smaller than the source domain[35], which made transfer learning a powerful tool in solving problems with insufficient training samples in various fields [20].

In 2021, Yang et al. facilitated human-virus protein-protein-interaction (PPI) data including HIV, Influenza, etc. and transferred these knowledge into human-SARS-CoV-2 PPI prediction[36]. Zhang et al. used general RNA-binding proteins(RBPs) as the initial training, and fine-tuned their model with species-specific RBP datasets to reach higher RBP identification accuracy[37].

When it comes to prediction of MS/MS spectra, pDeep2 facilitated transfer learning to reach higher efficiency on post-translational modification (PTM), regarding the limited PTM data sets [38]. In 2021, Chen et al. developed pDeepXL to predict MS/MS spectra of cross-linked peptide pairs. They leveraged a large dataset of MS/MS spectra of linear peptides, as well as datasets of cross-linked peptide pairs or cross-link spectrum matches that were 35-fold smaller than the linear peptides [16].

The number of identified MHC-associated peptides is quite limited, making it very hard to train a deep learning model from scratch. Transfer learning was a powerful technique to deal with insufficient training samples, [35] making it a good fit for MHC-associated peptides prediction

1.8 Contribution

In this thesis, we introduced that we fine-tuned a pre-published deep learning model, and reached better predictions for MS/MS spectra of MHC-Associated Peptides. This thesis is organized as following:

- In Chapter 2, we selected a pre-published deep learning model, and introduced how we fine-tuned this model.

- In Chapter 3 we described the datasets we selected, how we processed these datasets, and how we split train and test sets.

- In Chapter 4, we compared the accuracy of predicted spectra between the original model and the fine-tuned model.

- In Chapter 5, we discussed the reasons of the improvement after fine-tuning.

- In Chapter 6, we summarized this thesis, and discussed the potential work in future.

Chapter 2

Transfer Learning Methods

2.1 Selection of the Base Model

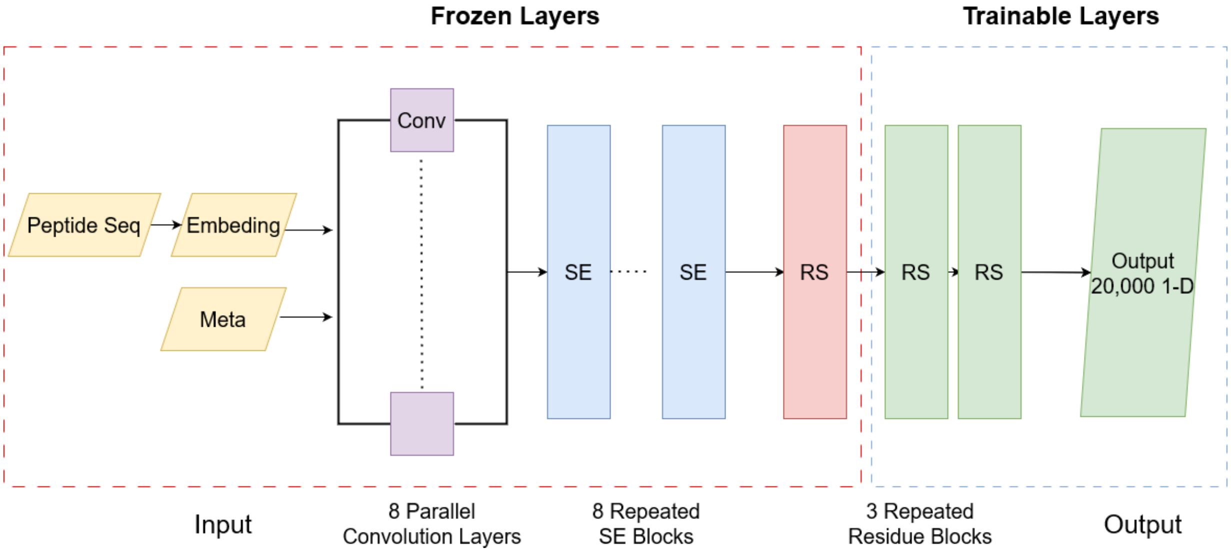


Figure 2.1: Architecture of the neural network. PredFull as the backbone model.

In our study, we selected PredFull[39] as our backbone model. It had the ability to predict the intensity for a/b/c/x/y/z ions, so we considered it to be with higher potential of reaching a better prediction accuracy. We remained the structure of the PredFull model, which facilitated the residual convolutional neural network [40]. As shown in Figure 2.1, it had two input layers, *embedding* and *meta*. Each peptide was converted into a 32 by 29 matrix for *embedding*, and a 3 by 30 *meta* for meta. The 2022-May PredFull model supports variable-length peptides, but considering the length of MHC-associated peptides, we limited the peptide length at 32 amino acids. In the *embedding* matrix, each amino acid was paired with the first dimension. In the second dimension, Pos 1-21 were used to represent the amino acid with one-hot encoding. All Leucine was considered as Isoleucine as they share the same mass. Pos 24 represents the monoisotopic mass of the amino acid after scaling. Pos 25 represents the peptide length. Pos 26-28 represents the modifications. In the *meta* matrix, peptide charge, peptide mass and Normalized Collision Energy(NCE) were stored.

A *pooling layer*, which was composed of eight parallel convolutional layers would receive the input data. These convolutional layers were 1-dimensional, with a constant filter size at 64 and their kernel sizes varied from 2 to 9. After merging, the result was passed into *8 sequential Squeeze-and-excitation blocks* (SE blocks) [41]. After that, there were *3 sequential residual blocks*[40]. Finally, the *output layer* decodes the previous layer's data into a vector with length of 20 000, matching the *Representation of MS/MS Spectrum* below.

2.2 Representation of MS/MS Spectrum

We followed the same method with PredFull to represent the MS/MS spectrum with an 1-D array[39]. A MS/MS spectrum was a list of peaks, and each peak had two properties, m/z and intensity, noted as (p, t). Firstly, the precursor peaks were removed from this list. Secondly, the peak with the highest intensity was selected. Its intensity was called t_{max}. Thirdly, the relative intensity was calculated. That is to say, for each peak, $r_i = t_i/t_{max}$. The peak with highest intensity would receive 1.0 as its relative intensity. Fourthly, squareroot would be applied to relative intensity for normalization, $s_i = \sqrt{r_i}$ Finally, the m/z range was restricted to 0-2000 and the bin width was set to 0.1 The index q of each peak was calculated by $q = \lfloor p * 10 \rfloor$. Therefore, it would result in an 1-D vector with 20 000 elements.

For example, given such spectrum.

```
M/Z     Intensity
187.108 6811.49
203.139 4380.94
560.379 14433.56
688.476 11370.93
```

It would be processed by such procedures:

1. There were no precursor ions in this spectrum.
2. The highest intensity $t_{max} = 14433.56$ was selected.
3. For the first peak, $(p_1, t_1) = (187.108, 6811.49)$, $r_1 = 6811.49/14433.56 = 0.4719$
4. Normalization: $s_1 = \sqrt{0.4719} = 0.6869$
5. Calculating the index: $q = 181.1 * 10 = 1871$

It would result in a very sparse 1-D vector with 20 000 elements. `v[1871]=0.6869`, `v[2031]=0.5509`, `v[5604]=1.0`, `v[6685]=0.8876`, and all other elements were filled with 0.

2.3 Fine-tuning and Hyper-parameters

2.3.1 Trainable Layers

Considering we were using a smaller training set for fine-tuning, only a small portion of the model was trainable. As shown in Figure 2.1, only the last two residual blocks and the output layer were set to trainable. All other layers remained frozen. They were considered to extract the features from the embedding and meta inputs.

2.3.2 Optimizer

Aligning with PredFull's training process, Adam Optimizer [42] was selected.

2.3.3 Loss Function

Cosine similarity was selected as the loss function, and Tensorflow built-in implementation was executed[26]. The safeguard of the division was different with the score function in section 2.4. In tensorflow, the cosine similarity was considered as 0 if one vector's norm was too close to zero.

2.3.4 Batch size

Regarding to the restriction of 16G VRAM, we set the batch size to 32. The peak memory usage was around 14G, showing that we've maximized the computational power of our hardware. During the 4-th cross-validation of the MHC-I Q-Exactive Model, the training process had crashed repeatedly, so the batch size was set to 16 for this run.

2.3.5 Epochs

For all three models, there were 100 epochs. When training the MHC-I Q-Exactive Model, the average time for one epoch was around 800s. For the MHC-I Fusion Lumos and Eclipse Tribrid Model, the time for each epoch reduced to 85s, due to the smaller training set. For the MHC Class II-associated peptides, each epoch cost 120s.

2.3.6 Learning Rate

For the MHC Class I-Associated Peptides Q-Exactive Instruments model, we selected 1e-5 as the learning rate. For the MHC-I Fusion Lumos and Eclipse Tribrid Model and the MHC Class II-associated Peptides Model, the learning rates were set to 7e-6. All of them were smaller than the learning rate for training the original PredFull model, which was 3e-4.

2.4 Score Function

Following the evaluation method by PredFull[39], cosine similarity between the predicted spectrum and real spectrum was selected as the score function, which was the criteria to reflect our model's improvement from the original model. As stated above in 2.2 , the predicted spectrum and real spectrum were represented as 1-D vectors, their raw intensities were replaced with relative intensities, and the precursor peaks were removed. It's worth noting that we choose the square root function to normalize the intensities, aligning with PredFull's evaluation. The cosine similarity of these two vectors was considered of the score of prediction.

Formally, when the experimental spectrum and the predicted spectrum are represented in 1-D vectors $\mathbf{u}$ and $\mathbf{v}$, after normalization and removing the precursor ions, the score function is defined as

$$S = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

There is a safeguard for computation that, if the norm for a vector is too small, 1e-16 would be filled as the divisor. However, as only the well-annotated experimental spectra were selected, and the intensities of the peaks were normalized in advance, this case was very rare to happen.

2.5 HardWare and Software Environment

Our server used two Intel Xeon Gold 6248 CPU @ 2.50GHz and one Nvidia Tesla V100 with 16G GPU Memory. Ubuntu Linux was installed on the server, and Docker 20.10.14 was available. We used NVIDIA TensorFlow Docker release (23.02-tf2-py3), which shiped CUDA (Version 11.6), TensorFlow (Version 2.8.0) and Python (3.8.10). We also bundled pyteomics 4.5 [43][44] and biopython 1.80 [45] in our Docker image to process spectra files.

We used MSConvert[46] [47] to process RAW files, and we used Comet (2022.01 rev. 2) [9], to perform the database search.

We used Jupyter Notebook, Matplotlib[48], Seaborn[49] and spectrum_utils [50] to do the data analysis and visualization.

Chapter 3

Data Source and Data Process

In this chapter, we described the datasets we selected, and introduced how we process these datasets, including the tools and parameters for database search

3.1 MS Data Selection

We're training three separate models, based on the classification of MHC-associated peptides and instruments.

3.1.1 MHC-I Q-Exactive Model

For this model, we used these following datasets:

- PXD004894 (MHC I-associated peptides) collected samples from melanoma with Q Exactive or the Q Exactive HF mass spectrometers [51]
- PXD008500 collected samples from four lymphoblastoid cell lines (LCL) which are HLA-B*27:05-positive [52] .
- PXD022020 collected samples from Jurket cells[53] .
- PXD026702 used Q-Exactive HF to study pluripotency-associated MHC I peptides (paMAPs) [31].

Inspired by previous research [54], we only use PXD004894 to do the fine-tuning. The other three datasets were only used to evaluate the performance of the fine-tuned Q-Exactive model.

Database ID	PSMs
PXD004894 (MHC-I)	671,551
PXD008500	35,399
PXD022020	19,074
PXD026702	26,002
PXD024965	2,464
PXD011723	48,729
PXD024562	16,539

Table 3.1: The number of PSMs in each MHC-I peptide datasets

3.1.2 MHC-I Fusion Lumos and Eclipse Tribrid Model

For this, we used these three datasets for training and evaluation

- PXD024965 uses Orbitrap Fusion Lumos on breast cancer cell lines [55]
- PXD011723 used Orbitrap Fusion Lumos on Jurket cells [56]
- PXD024562 used Orbitrap Eclipse Tribrid on Loucy and A375 cell lines.[57]

3.1.3 MHC-II Model

MHC II-associated peptides in PXD004894 [51] were used to train and evaluated the MHC-II model.

3.2 File format conversion

The MS/MS spectrum were provided in RAW format. We facilitated MSConvert, a tool under ProteoWizard[46] [47] to convert RAW files into mgf format[58], which was a plain text format, and could be fed into various softwares and scripts.

Converting a RAW file into other formats with MSConvert needed to import a closed-source DLL file. As we were running it on our server which ran Ubuntu Linux, we followed MSConvert's guidance to run with wine. The dependency, wine, was packed in MSConvert's official docker image.

```
    docker run -it --rm -e WINEDEBUG=-all \
-v /data/zhenboli/library:/data_in \
-v /data/zhenboli/result:/data_out \
chambm/pwiz-skyline-i-agree-to-the-vendor-licenses \
wine msconvert \
/data_in/raw_spectrum_file.raw \
--filter "peakPicking true 2-"  \
--filter "threshold absolute 0 most-intense" \
-o /data_out  --mgf
```

3.3 Database Search

Considering Comet has a native Linux release, and it could facility the multi-core CPU on our server, we used Comet (2022.01 rev. 2) [9] for the database search. We search against the UniProt database [59] (Human 80,851 entries, Dec 2022). We selected HCD instruments and set the fragment ion tolerance to 0.02 Da. The peptide mass tolerance varies by dataset (7 to 20 ppm). We disabled modifications. We set no specific enzyme and allowed no missed cleavage. We performed FDR control at 1% level on Comet results. For MHC I-associated peptides, we filtered out peptides with 13 or more amino acids. The result of our database search process is shown in Table 3.1

The search command was:

```
    ./comet.linux.exe -Duniprot_human.fasta raw_spectrum_file.mgf
```

3.4 Dataset Split

Using the same backbone model, we trained three separate models for to reach the best performance.

Partition	Patients ID	PSMs
P1	M15	225,116
P2	M16	92,708
P3	M3,M20,M21,M28,M30,M35,M38	95,882
P4	M24,M26,M27,M29,M34,M36,M40	96,247
P5	M4,M12,M25,M33,M39,M42	93,671
Independent Test Set	M5,M8,M41	67,927

Table 3.2: Grouping PXD004894 into 5 Partitions and an Independent Test Set by Patients

3.4.1 Q-Exactive Family

We solely used MHC Class-I associated peptides in PXD004894[51] for training the Q-Exactive model with a five-fold cross validation, and other Q-Exactive datasets were only used for evaluation. Following the approach from Liu et al. [60], we selected roughly 90% PSMs of PXD004894 as the cross-validation set. The remaining 10% PSMs were selected as the independent test set and used to benchmark the prediction of our model. PXD004894 was composed of data from 25 patients and all the observations of the same patient should be gathered into the same partition. Hence, making equal-size partitions was a multiway number partitioning problem, proved to be NP-hard[61]. We applied a greedy algorithm to approximately divide the cross-validation set into five partitions evenly. The result was shown in Table 3.2.

3.4.2 Fusion Lumos and Eclipse Tribrid

We used all three PXD024965[55], PXD011723[56] and PXD024562[57] to train and evaluate our Fusion Lumos and Eclipse Tribrid model. For each of them, we firstly grouped all PSMs with the bin of 0.01 Da. Considering that the same peptide may appear in the dataset with slightly different mass, grouping the PSMs into bags could avoid the same peptide appears in both training and testing dataset. The numbers of the PSMs and bags were shown in Table 3.3

We randomly selected 10% bags of each dataset as the independent test and split the remaining 90% bags into five partitions. That is to say, each partition would include 18% bags. Similarly to the Q-Exactive Model, four partitions were used as the training set, and the remaining partition was used as the validation set for this 5-fold cross-validation.

ID	PXD024965	PXD011723	PXD024562
PSMs	2,464	48,729	16,539
Bags	1,410	4,624	5,920

Table 3.3: The number of PSMs and Bags of Fusion Lumos and Eclipse Tribrid datasets

3.4.3 MHC II-Associated Peptides

We used the MAP Class II part of PXD004894 to train and evaluate the MHC Class II model. Similar to above, we grouped the 82692 PSMs by 0.01 Da, resulting in 12278 bags. We randomly selected 10% bags as the independent test set, and the remaining 90% bags were used for a three-fold cross validation.

Chapter 4

Result

4.1 Q-Exactive Model

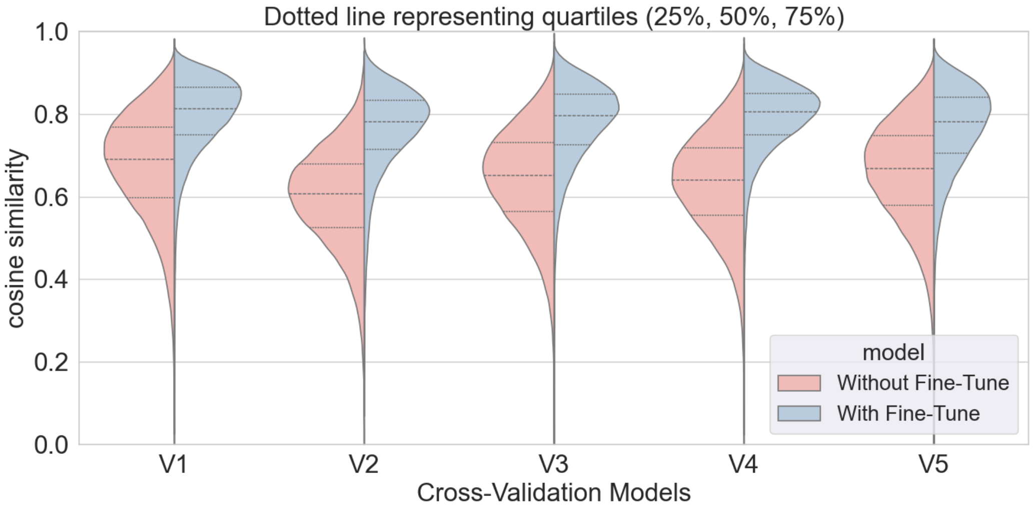


Figure 4.1: Five-fold cross-validation on PXD004894. The dotted lines represent the quartile (25%,50%,75%) of the distribution. V1: partitions P2,P3,P4,P5 were used as the training set, and partition P1 were used as the validation set.

4.1.1 Performance on Q-Exactive family datasets with 5-fold Cross validation

In this part, a five-fold cross-validation was performed to compare our fine-tuned model with the original PredFull model. As stated in Table 3.2, the PXD004894 dataset was grouped into partitions by patients. As shown in Figure 4.1, the quartile increased from (0.525,0.606,0.679) to (0.714,0.782,0.834), showing a 35% increase. The lower quartile (25%) of the fine-tuned models was even higher than the medium(50%) of the PredFull model. The average of cosine similarity between predicted spectra and real spectra increased from 0.6477 to 0.7820. These results suggested that fine-tuned model outperformed the PredFull model.

4.1.2 Performance on New Q-Exactive Family Datasets

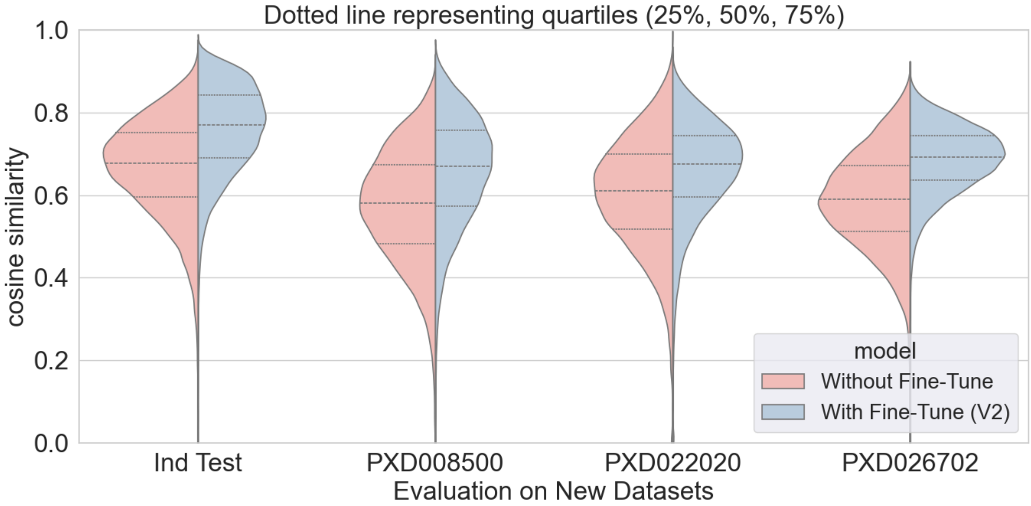


Figure 4.2: The dotted lines represent the quartile (25%,50%,75%) of the distribution. The violin plot showed the performance on new datasets tested on the V2 model.

To test the ability of our fine-tuned model for predicting the MS/MS spectrum of new datasets, we tested its performance on three new Q-Exactive datasets that didn't appear in the cross-validation. We also tested this fine-tuned model against the independent test set, which contained around 10% PSMs of PXD004894. Among the five cross-fold models in Section 4.1.1, the V2 model was selected for evaluation.

Figure 4.2 showed that on the Independent Test Set, the average of cosine similarity increased from 0.6658 to 0.7577(+13.8%). The lower quartile of the fine-tuned model was 0.6895, which was higher than the median of the original model, 0.6777. On PXD008500, the average cosine similarity increased from 0.5729 to 0.6573(+14.7%), and the median increased from 0.5801 to 0.6705(+15.58%). On PXD022020, the average increased from 0.6024 to 0.6601(+9.6%). And on PXD026702, the average increased from 0.5892 to 0.6853(+16.3%). The fine-tuned model showed higher prediction accuracy on all four new datasets compared to the original model.

4.1.3 Pairwise Performance Evaluation on Q-Exactive Family Datasets

We also compared our re-trained model with PredFull pair-wisely. That is to say, for each observation in the test dataset, we used the PredFull model (Without Fine-Tune) and the Fine-Tuned model to predict the spectra for the given peptide sequence. Using the evaluation method mentioned in Section 2.4, we calculated the square of cosine similarity between the predicted spectrum and the experimental spectrum. We used scatter plots to represent the prediction scores of the two models. As shown in Figure 4.3, the fine-tuned model performed better for most peptide sequences in all four datasets.

4.1.4 Case Study of Predicted MS/MS Spectrum Before and After Fine-Tuning

In this section, we had a deeper look at examples of individual peptides from the independent test set. We picked four sequences as examples to represent improved, stalled and worsened predictions after the fine-tuning.

Peptide TASEMILVL was from the independent test set. The information of immonium ions might benefit peptide identification, including reducing the size of the search space for improving the speed in large-scale database search[62]. The original model didn't output immonium ion peaks of TASEMILVL (left-most part in Figure 4.4a), while the fine-tuned model predicted the immonium ion peaks successfully (Figure 4.4b). The original model ignored b5 peak in its output, and overestimated the intensities of y1 and b7 peaks. The fine-tuned model predicted these three peaks with much more accurate intensities.

Peptide LPQLTGAENVL was another example from the independent set, showing the improvement of the fine-tuned model. In Figure 4.4c, the original model ignored

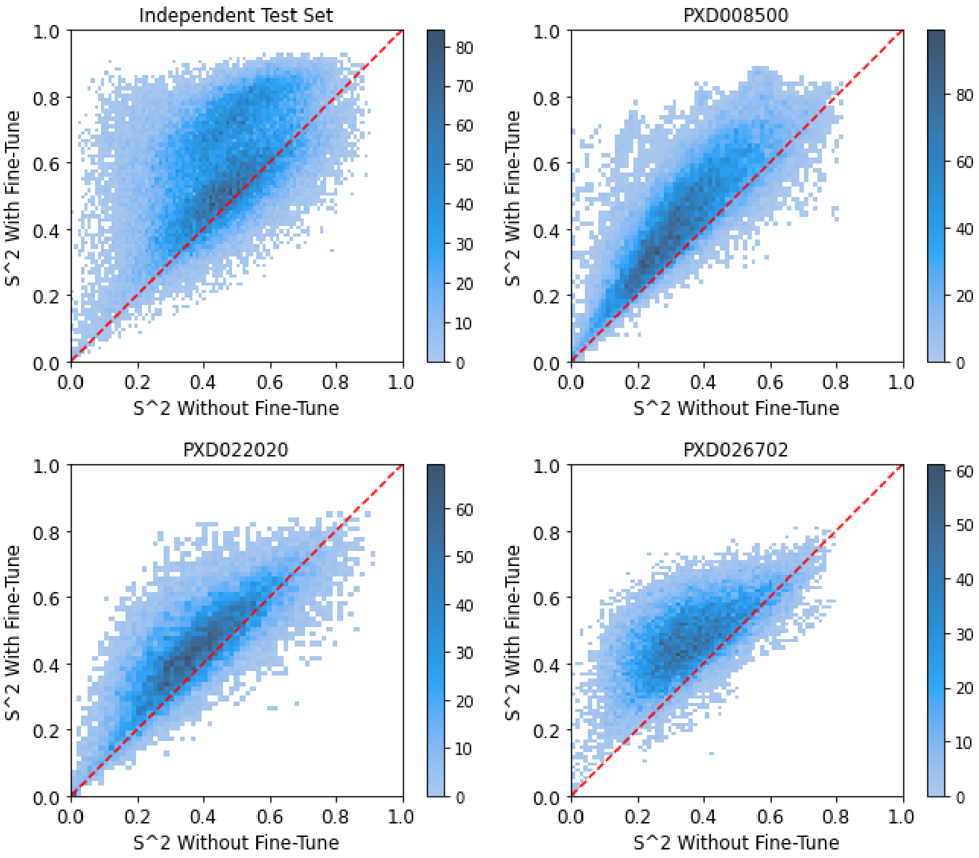
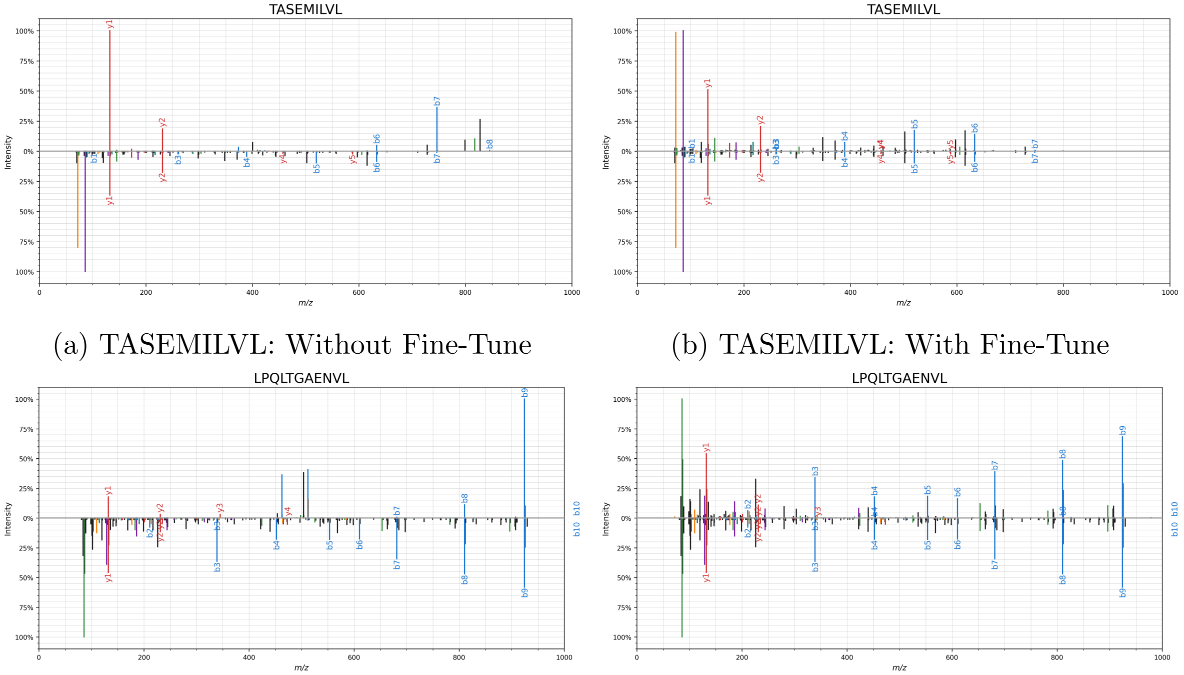


Figure 4.3: Pairwise comparison of each observation in the four new datasets. S^2 was the square of cosine similarity between the real spectrum and the predicted spectrum. For each observation, X-axis represented S^2 for the original model before fine-tuning, and Y-axis represented the fine-tuned model.

immonium ion peaks, while the fine-tuned model predicted the immonium ion successfully (Figure 4.4d). The fine-tuned model displayed b3, b5, b6, b7 which were missed by the original model. Also, the fine-tuned model predicted the intensities of y1, b8, b9 peaks more accurately.

Transfer learning didn't result in better prediction accuracy for peptide TSERTVLRY. Before the fine-tuning, the cosine similarity between the predicted spectrum and the real spectrum was 0.7282, which was already an accurate prediction (Figure 4.5a) After fine-tuning, the cosine similarity slightly dropped to 0.7182.It suggested that fine-tuning had very little effect on the model's prediction of peptide TSERTVLRY (Figure 4.5b).



(a) TASEMILVL: Without Fine-Tune

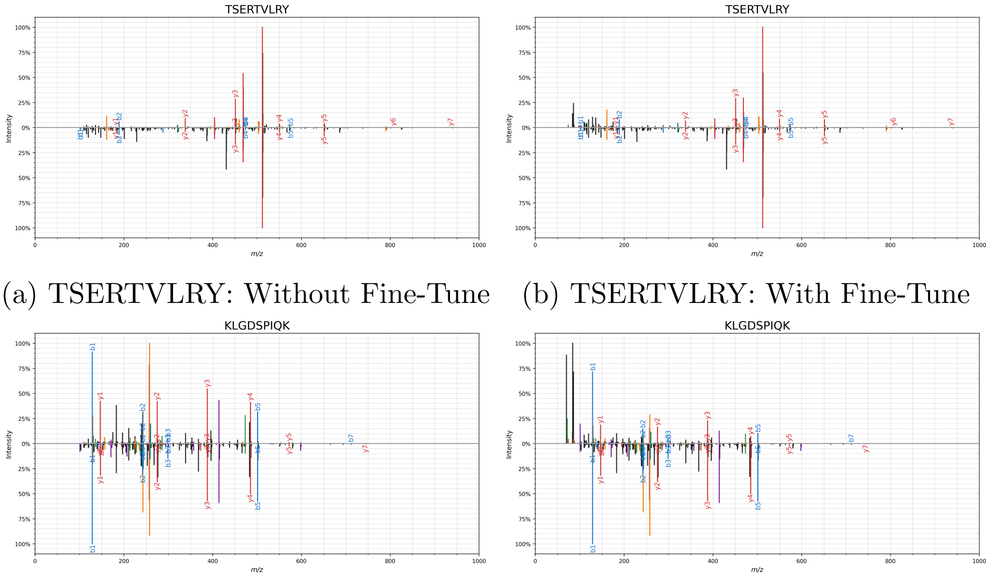
(b) TASEMILVL: With Fine-Tune

(c) LPQLTGAENVL: W/O Fine-Tune

(d) LPQLTGAENVL: With Fine-Tune

Figure 4.4: Example of predictions improved after fine-tuning. In each subfigure, the upper part was the predicted spectrum and lower part was the real mass spectrum.

KLGDSPIQK was an example of decreased performance. As shown in Figure 4.5c , the original model predicted the spectrum with a high accuracy and all the immonium ion peaks were predicted successfully. After fine-tuning(Figure 4.5d), two fake immonium ion peaks were generated, and the cosine similarity dropped from 0.8338 to 0.6308.



(a) TSERTVLRY: Without Fine-Tune (b) TSERTVLRY: With Fine-Tune

(c) KLGDSPIQK: Without Fine-Tune (d) KLGDSPIQK: With Fine-Tune

Figure 4.5: Example of predictions with no improvements after fine-tuning. In each subfigure, the upper part was the predicted spectrum and lower part was the real mass spectrum.

4.2 Prediction Performance on Fusion Lumos and Eclipse Tribrid Datasets

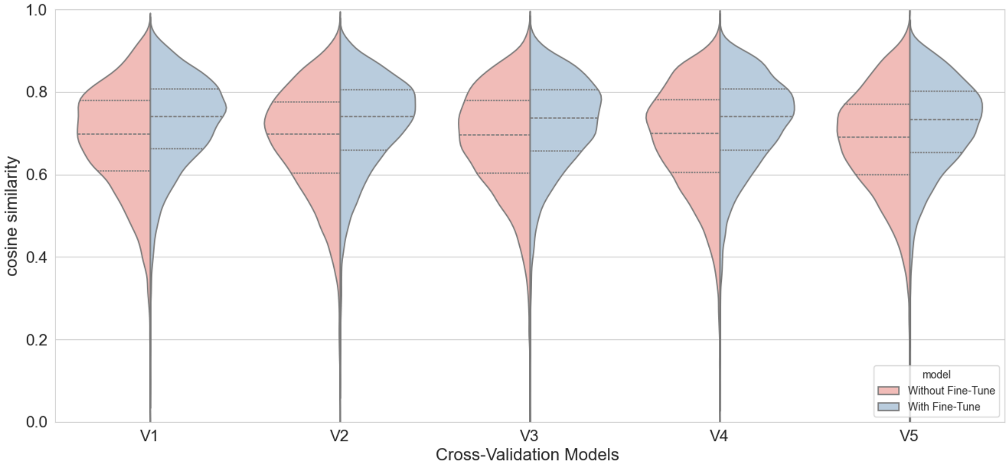


Figure 4.6: Five-fold cross-validation on Fusion Lumos and Eclipse Tribrid datasets. The dotted lines represent the quartile (25%,50%,75%) of the distribution. V1 means we used partitions P2,P3,P4,P5 as the training set, and partition P1 as the validation set.

Similarly to the Q-Exactive model, a five-fold cross validation was applied on three these datasets, PXD011723, PXD024562 and PXD024965. As shown in Figure 4.6, the average cosine similarity between the predicted spectrum and the real spectrum increased from 0.6828 to 0.7237(+6%). The the quartile (25%,50%,75%) increased from (0.6039, 0.6959, 0.7772) to (0.6575, 0.7375, 0.8050). It indicated that the fine-tuned model had better prediction in the cross-validation.

We also evaluated the performance on the independent test set, composed of roughly 10% PSMs which never appeared in the cross-validation process. As figure 4.7 showed, the average cosine similarity increased from 0.6820 to 0.7259 (+6%). The quartile (25%,50%,75%) increased from (0.5995, 0.6949, 0.7806) to (0.6591, 0.7388, 0.8076). It suggested that the fine-tuned model was capable of providing more accurate predictions on unseen datasets compared to the original model.

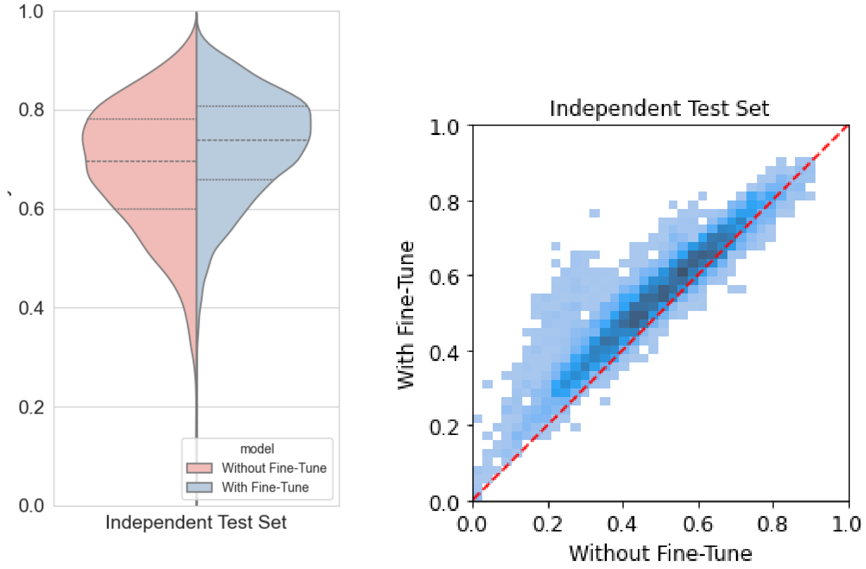


Figure 4.7: Evaluation on Independent Test Set of Fusion Lumos and Eclipse Tribrid

4.3 Prediction Performance on MHC II-Associated Peptides

We performed a three-fold cross validation on the MHC II-associated peptides. Shown in Figure 4.8, After fine-tuning, the average of cosine similarity between predicted spectrum and real spectrum had an 18.7% increase (from 0.5937 to 0.7051). The quartile (25%,50%,75%) also increased from (0.4966, 0.6158, 0.7120) to (0.6434, 0.7301, 0.7945).

We evaluated the fine-tuned model's performance on the independent test set on the V1 model. As shown in Fig 4.9, a high quality performance was observed. The average of cosine similarity had a 20.2% increase, from 0.5828 to 0.7003. The quartile (25%,50%,75%) increased from (0.4799, 0.6028, 0.7033) to (0.6413, 0.7263, 0.7883).

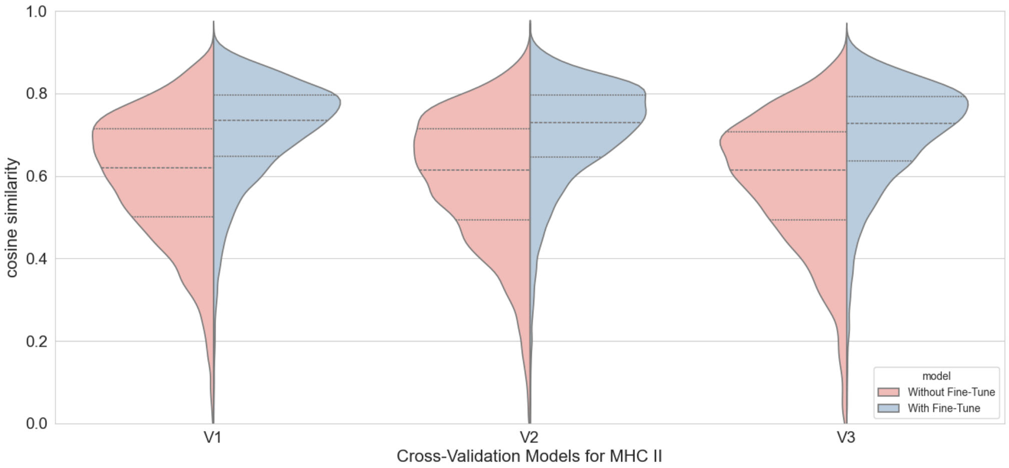


Figure 4.8: Three-fold cross-validation on MHC II-associated peptides from PXD004894. The dotted lines represent the quartile (25%,50%,75%) of the distribution. V1 means we used partitions P2,P3 as the training set, and partition P1 as the validation set.

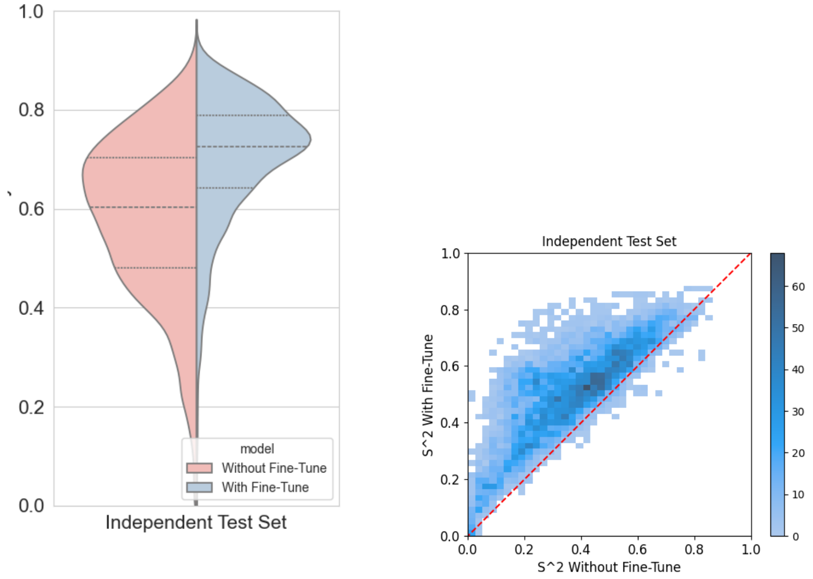


Figure 4.9: Evaluation on Independent Test Set of MAP II-associated peptides

Chapter 5

Factor of Improvements Analysis

In this chapter, we looked deeper into the improvements of the prediction. There is a famous idiom in the early ages of computer science: Garbage In, Garbage Out.[63] When it comes to training a machine model, the quality of training data is crucial to the quality of the model. The bias present in the training data would result in the constraints of the model's prediction.

5.1 Tryptic peptides were more likely to be improved by fine-tuning

The majority of published MS/MS spectrum predictors were trained with tryptic peptide spectral libraries. However, most MHC-associated peptides were non-tryptic peptides. For example, Prosit[29] used PXD004732[25] and PXD010595[29] to train their model, which were both tryptic peptide datasets. Guan et al. [28], used another tryptic peptide dataset, MassIVE-KB [64]. Our base model, PredFull, used NIST HCD library[24], NIST Synthetic HCD library, MassIVE Human HCD library , ProteomeTools synthetic HCD library [25]. All these datasets were tryptic.

Dataset **NIST HCD** consisted of 605,927 unique peptides. 290519 of them were ended with K, and 266563 were ended with R.

NIST Synthetic HCD library consisted of 188,804 peptides and more than 90% were ended with R or K.

MassIVE v2 consisted 2,512,045 unique peptides, 40% ended with K and 38% ended with R

ProteomeTools synthetic HCD library consisted of 391,273 unique peptides and more than 96% were ended with R or K.

In total, PredFull was trained with tryptic peptides and more than 90% of its input peptides were ended with K or R. In the independent test set, there were 67,927 PSMs with 22,816 unique peptides, of which 32% ended with K, and 4% ended with R. That is to say, only 36% peptides were tryptic peptides.

5.1.1 The original model's performance dropped with non-tryptic models

Firstly, we compared the original PredFull model's performance on tryptic peptides with non-tryptic peptides. As shown in Fig 5.1a, typtic peptides had better predictions. This result was validated through an one side t-test, yielding p=0.0 (too close to zero).

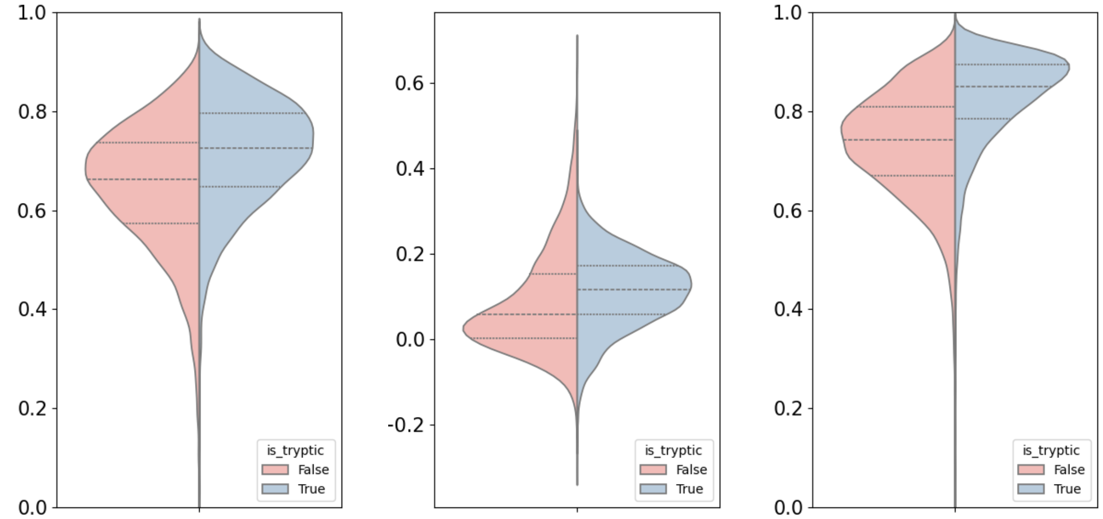
5.1.2 Fine-Tuning Improved Performance on both Tryptic and Non-Tryptic Peptides

Given that the original model exhibited a higher likelihood of inaccurately predicting non-tryptic peptides, we focused on if the fine-tuned model had improved performance on tryptic and non-tryptic peptides.

Among 22,816 peptides in the independent test set, 15,437 peptides increased at least 5%, 12,426 increased at least 10%, 5,800 increased 25% and 1622 peptides increased more than 50%. In these 1622 peptides, 35% ended with L and only 14.5% end with K.

We defined the *changed score* as the fine-tuned model's prediction score minus the original model's prediction score. We performed a T-test on all the peptides' improved scores (shown in Fig 5.1b). It showed that tryptic peptides were more likely to be improved with the fine-tuning, yielding p=1.8e-65.

After fine-tuning, the prediction quality for tryptic peptides was better than non-tryptic peptides (Fig 5.1c)



(a) Before Fine-Tuning

(b) With vs. W/O

(c) After Fine-Tuning.

Figure 5.1: Distribution of prediction score in the Independent Test Set. Score function was the cross similarity of the predicted spectrum and the experimental(real) spectrum

5.2 Effect of amino acids with possitive charges(K,R,H)

There are three positive charged amino acid: Lysine (Lys, K), Arginine (Arg, R) and Histidine (His, H) . We conducted an in-depth analysis of their impact on the model's performance.

In the datasets used for training the original PredFull model (see section 5.1.1), more than 80% of the peptides had 1-2 positive amino acids. As they were tryptic datasets, less than 0.5% of the peptides contained no amino acids with positive charges.

In the independent test set, the distribution of positively charged amino acids were similar except roughly 10% peptides contained no K,R or H.

A peptide with no Lys, Arg or His would definitely be a non-tryptic peptide. Figure 5.2 showed that the original model and the fine-tuned model had similar patterns of the relationship between number of positively charged amino acids and the quality of prediction. The scores of all predictions were significantly higher after fine-tuning.

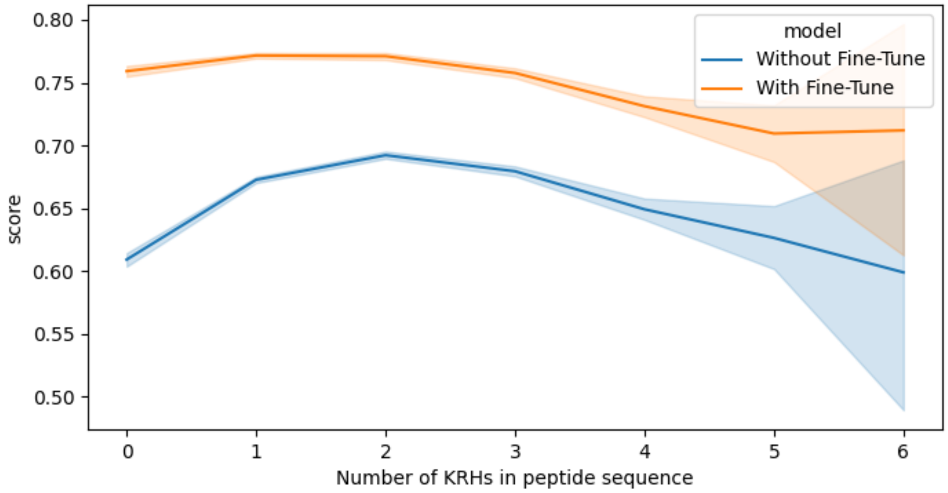


Figure 5.2: The relationship between the number of Lys, Arg and His in the peptide and the models' prediction score

5.3 Effect of Peptide Charges

In the Independent Test Set, 62% peptides had 2+ charge, 25% peptides had 3+ charge and only 13% peptides had 1+ charge.

As shown in Figure 5.3, the model's prediction on peptides with all charges were improved by the fine-tuning.

5.4 Effect of Proline

Proline (Pro; P) is a unique amino acid with a cyclinc structure. In both the PredFull's training sets and the independent data set, there were about 40-50% peptides containing at least one Proline.

Figure 5.4a showed that if one or more Proline appeared in the peptide, the original model tended to provide worse predictions. An one-side t-test confirmed this conclusion, yielding p=7.3e-27.

We tested the same dataset on the fine-tuned model and compared the prediction score between the fine-tuned and the original model. As before, *changed score* was defined as the

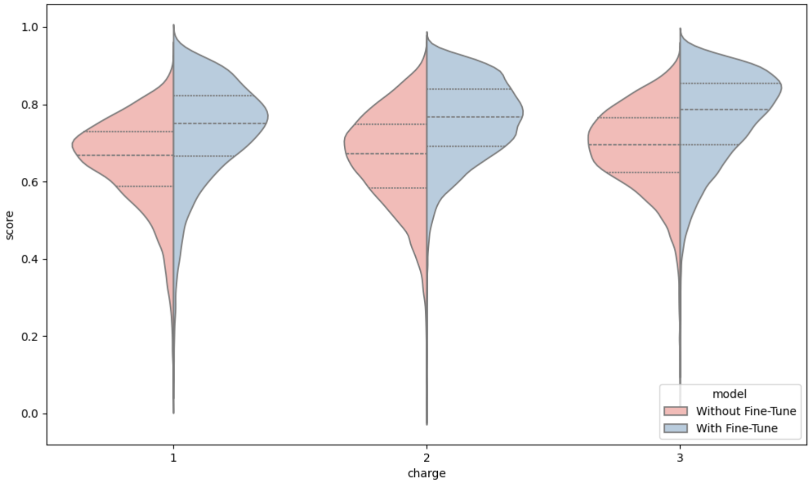


Figure 5.3: Prediction Score for different peptides' charges. Left: Without Fine-Tune. Right: With Fine-Tune

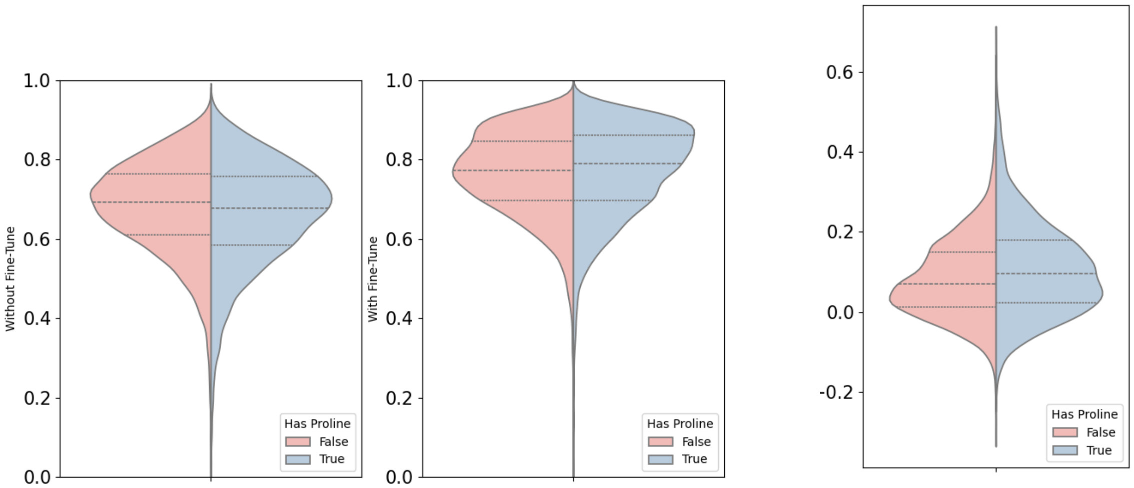
prediction score of the fine-tuned model minus the original model's. As shown in Figure 5.4b , prediction scores of all peptides were higher due to fine-tuning, and peptides with Proline tended to receive more significant improvements. An one-tailed T-Test supported this conclusion,yielding p=4.4e-65.

5.5 Number of Annotated peaks in the Experimental Spectrum

5.5.1 Definition of Less Peak Peptides

we used spectrum_utils[50] to annotate the peaks in both the real spectra and the predicted spectra. We considered primary peptide fragments (abcxyz) and allowed 0.02 Da as fragment tolerance.

In the independent test set, among all 67,927 PSMs, each real spectrum had 30.05 annotated peaks on average. The quartile (25%, 50%, 75%) of annotated peaks were 24, 30, 36. Therefore, if a peptide's number of peaks was in the lower 25% bracket(less than 25 annotated peaks), we considered this peptide as a *less peak* peptide.



(a) Model's Prediction Score. Left: Without Fine-Tune. Right: With Fine-Tune

(b) Improvement (changed score) after Fine-Tuning

Figure 5.4: Effect of Proline.

5.5.2 Performance with Less Peak Peptides

When the real spectra had a small number of annotated peaks, the original model's prediction quality dropped. The original model's average prediction score on *less peak* peptide was 0.645, which was lower than its performance on other peptides 0.672 (Fig 5.5a). After fine tuning, this gap was closed. The fine-tuned model's average prediction score is 0.764 for *less peak* peptides, and 0.738 for other peptides(Fig 5.5b).

5.6 Summary

In this chapter, we found that the fine-tuning had such effects:

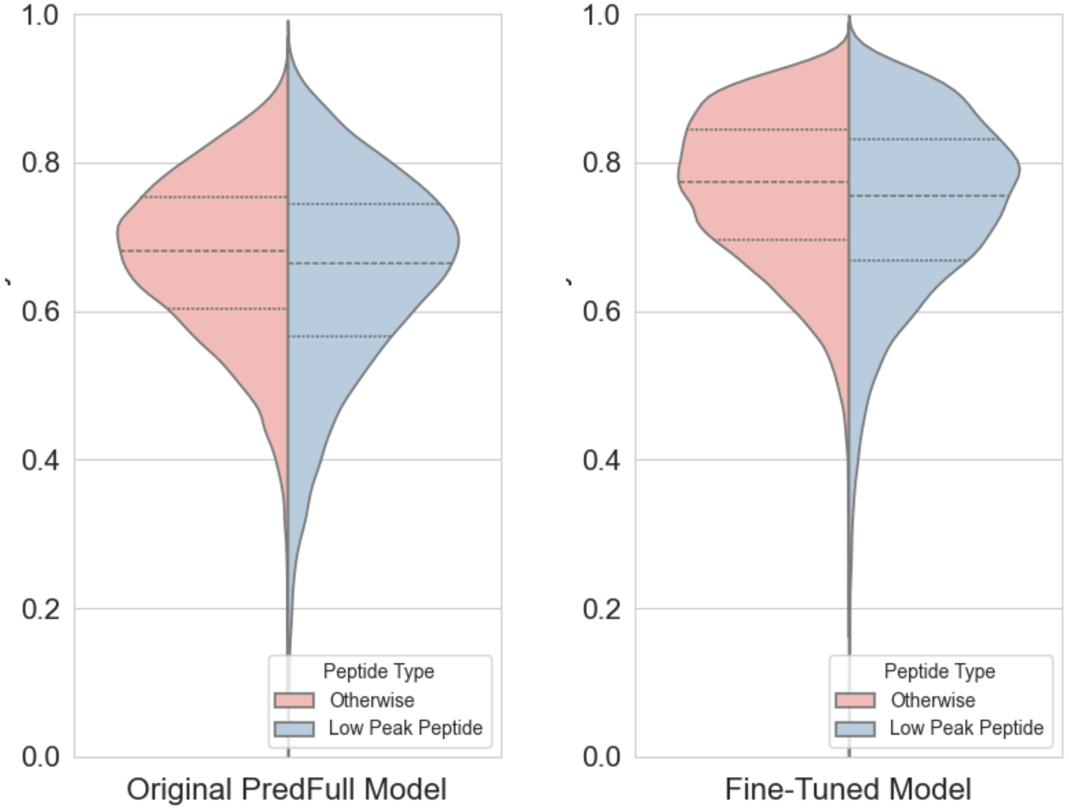
- Improving performance of both tryptic and non-tryptic peptides. Tryptic peptides received more significant enhancements.

- Improving performance of peptides with various number of positively charged amino acids (Lys, Arg and His).

- Improving performance of peptides with different charges.

- Improving performance of peptides with Proline.

- Improving performance of peptides with low numbers of annotated peaks (*less peak* peptides).



(a) Original model's distribution of prediction score in the Independent Test Set.

(b) Fine-tuned model's distribution of prediction score in the Independent Test Set.

Figure 5.5: Low-peaks means the real spectrum had less than 25 annotated peaks

Chapter 6

Conclusion

In this thesis, we picked up a pre-published deep learning model for prediction of MS/MS spectrum, applied transfer learning on it and reached better accuracy when predicting MAC associated peptides.

In the beginning, as we found no public spectral libraries for MHC-associated peptides, we applied database search to build them. They were used for transfer learning and evaluation of the fine-tuned model. Next, we applied transfer learning on a pre-published model. Due to the different properties of the peptides, we trained three models. Considering the size of the training datasets, we reduced learning rates to reach the best loss on validation set.

During the evaluation, we observed all these three models had significant improvement after fine-tuning. We also applied further investigation on the MHC Class I with Q-Exactive instruments. It could provide better predictions on datasets who didn't appear in training or validation set as well. In addition, we found that the fine-tuned model was more likely to have better predictions for immonium ions. Lastly, we showed that the improvement might came from the improved coverage for non-tryptic peptides.

In the future, we propose these improvements

Unite the three models into one. In our research, we found that the improvement in MHC Class I with Q-Exactive instruments would decrease the performance in MHC Class I with other instruments. By modifying the input layer of the model, and to add more meta data in the model input, including instruments and types of MHC associated peptides, it may be possible to reach good prediction accuracy with a single model.

Provide better training dataset. The quality of input dataset was crucial to deep learning. As the data of MHC associated peptides were limited, the power of the database search engine was more crucial. We selected Comet for several reasons: (1) Comet is across-platform software, so we could integrate it with other workflows on our workstation, which runs Ubuntu Linux. (2) Comet has good multi-thread supporting, which could reach the best efficiency on our workstation. (3) Comet is an open-source software and it is free of charge. However, as Parker et al. pointed out, Comet might not be the best search engine. In their research, they found that Peaks performed the best when searching for MHC Class I associated peptides. MSGF+ also performed better than Comet [65]. Switching the search engine may enlarge the spectral libraries, thus improving the performance of the fine-tuned model.

References

[1] Safa Al-Amrani, Zaaima Al-Jabri, Adhari Al-Zaabi, Jalila Alshekaili, and Murtadha Al-Khabori. Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, 12(5):57–69, September 2021.

[2] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. Proteomics: Technologies and Their Applications. *Journal of Chromatographic Science*, 55(2):182–196, February 2017.

[3] Yaoyang Zhang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews*, 113(4):2343–2394, April 2013. Publisher: American Chemical Society.

[4] Patricia Hernandez, Markus Müller, and Ron D. Appel. Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrometry Reviews*, 25(2):235–254, 2006. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/mas.20068.

[5] precursor ion. *IUPAC Compendium of Chemical Terminology, 3rd ed.*, 2019.

[6] Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. 5(9):699–711. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Reviews Publisher: Nature Publishing Group.

[7] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, November 1994. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[8] Jimmy K. Eng, Brian C. Searle, Karl R. Clauser, and David L. Tabb. A Face in the Crowd: Recognizing Peptides Through Database Search*. *Molecular & Cellular Proteomics*, 10(11):R111.009522, November 2011.

[9] Jimmy K. Eng, Tahmina A. Jahan, and Michael R. Hoopmann. Comet: An open-source MS/MS sequence database search tool. 13(1):22–24. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201200439.

[10] Robertson Craig and Ronald C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*, 20(9):1466–1467, June 2004.

[11] R. Craig, J. C. Cortens, D. Fenyo, and R. C. Beavis. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *Journal of Proteome Research*, 5(8):1843–1849, August 2006. Publisher: American Chemical Society.

[12] David Fenyö and Ronald C. Beavis. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Analytical Chemistry*, 75(4):768–774, February 2003.

[13] Henry Lam, Eric W. Deutsch, James S. Eddes, Jimmy K. Eng, Nichole King, Stephen E. Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *PROTEOMICS*, 7(5):655–667, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.200600625.

[14] Eric W. Deutsch, Luis Mendoza, David Shteynberg, Terry Farrah, Henry Lam, Natalie Tasman, Zhi Sun, Erik Nilsson, Brian Pratt, Bryan Prazen, Jimmy K. Eng, Daniel B. Martin, Alexey Nesvizhskii, and Ruedi Aebersold. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics*, 10(6):1150–1159, March 2010.

[15] Sven Degroeve and Lennart Martens. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, 09 2013.

[16] Zhen-Lin Chen, Peng-Zhi Mao, Wen-Feng Zeng, Hao Chi, and Si-Min He. pDeepXL: MS/MS spectrum prediction for cross-linked peptide pairs by deep learning. 20(5):2570–2582. Publisher: American Chemical Society.

[17] Zhongqi Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. 76(14):3908–3922. Publisher: American Chemical Society.

[18] Randy J. Arnold, Narmada Jayasankar, Divya Aggarwal, Haixu Tang, and Predrag Radivojac. A MACHINE LEARNING APPROACH TO PREDICTING PEPTIDE FRAGMENTATION SPECTRA. In *Biocomputing 2006*, pages 219–230. WORLD SCIENTIFIC.

[19] Sujun Li, Randy J. Arnold, Haixu Tang, and Predrag Radivojac. On the accuracy and limits of peptide fragmentation spectrum prediction. 83(3):790–796.

[20] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. 8(1):53.

[21] Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Computer Methods and Programs in Biomedicine*, 213:106504, January 2022.

[22] Henning Hermjakob and Rolf Apweiler. The proteomics identifications database (pride) and the proteomexchange consortium: making proteomics data accessible. *Expert review of proteomics*, 3(1):1–3, 2006.

[23] Eric W Deutsch, Nuno Bandeira, Yasset Perez-Riverol, Vagisha Sharma, Jeremy J Carver, Luis Mendoza, Deepti J Kundu, Shengbo Wang, Chakradhar Bandla, Selvakumar Kamatchinathan, Suresh Hewapathirana, Benjamin S Pullman, Julie Wertz, Zhi Sun, Shin Kawano, Shujiro Okuda, Yu Watanabe, Brendan MacLean, Michael J MacCoss, Yunping Zhu, Yasushi Ishihama, and Juan Antonio Vizcaíno. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Research*, 51(D1):D1539–D1548, 11 2022.

[24] Sergey L. Sheetlin, Guanghui Wang, Dmitrii V. Tchekhovskoi, Zheng Zhang, and Stephen E. Stein. Filtering and optimization of peptide tandem mass spectral libraries. 68th ASMS Conference on Mass Spectrometry Allied Topics, 2020.

[25] Daniel P. Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J. Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W. Deutsch, Ruedi Aebersold, Robert L. Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer,

and Bernhard Kuster. Building ProteomeTools based on a complete synthetic human proteome. 14(3):259–262. Number: 3 Publisher: Nature Publishing Group.

[26] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Shenheng Guan, Michael F. Moran, and Bin Ma. Prediction of LC-MS/MS properties of peptides from sequence by deep learning*[s]. 18(10):2099–2107.

[29] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. 16(6):509–518. Number: 6 Publisher: Nature Publishing Group.

[30] Massimo Andreatta, Annalisa Nicastri, Xu Peng, Gemma Hancock, Lucy Dorrell, Nicola Ternette, and Morten Nielsen. MS-rescue: A computational pipeline to increase the quality and yield of immunopeptidomics experiments. 19(4):1800357. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201800357.

[31] Anca Apavaloaei, Leslie Hesnard, Marie-Pierre Hardy, Basma Benabdallah, Gregory Ehx, Catherine Thériault, Jean-Philippe Laverdure, Chantal Durette, Joël Lanoix,

Mathieu Courcelles, Nandita Noronha, Kapil Dev Chauhan, Sébastien Lemieux, Christian Beauséjour, Mick Bhatia, Pierre Thibault, and Claude Perreault. Induced pluripotent stem cells display a distinct set of MHC i-associated peptides shared by human cancers. 40(7). Publisher: Elsevier.

[32] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2016.

[33] Xuezhi Xie, Yuanyuan Han, and Kaizhong Zhang. MHCherryPan. a novel model to predict the binding affinity of pan-specific class i HLA-peptide. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 548–554.

[34] Etienne Caron, Daniel J. Kowalewski, Ching Chiek Koh, Theo Sturm, Heiko Schuster, and Ruedi Aebersold. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. 14(12):3105–3117.

[35] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[36] Xiaodi Yang, Shiping Yang, Xianyi Lian, Stefan Wuchty, and Ziding Zhang. Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. 37(24):4771–4778.

[37] Jun Zhang, Ke Yan, Qingcai Chen, and Bin Liu. PreRBP-TL: prediction of species-specific RNA-binding proteins based on transfer learning. 38(8):2135–2143.

[38] Wen-Feng Zeng, Xie-Xuan Zhou, Wen-Jing Zhou, Hao Chi, Jianfeng Zhan, and Si-Min He. MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. 91(15):9724–9731. Publisher: American Chemical Society.

[39] Kaiyuan Liu, Sujun Li, Lei Wang, Yuzhen Ye, and Haixu Tang. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Analytical chemistry*, 92(6):4275–4283, 2020.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[41] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. 42(8):2011–2023. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[43] Anton A. Goloborodko, Lev I. Levitsky, Mark V. Ivanov, and Mikhail V. Gorshkov. Pyteomics—a python framework for exploratory data analysis and rapid software prototyping in proteomics. *Journal of the American Society for Mass Spectrometry*, 24(2):301–304, 2013. PMID: 23292976.

[44] Lev I. Levitsky, Joshua A. Klein, Mark V. Ivanov, and Mikhail V. Gorshkov. Pyteomics 4.0: Five years of development of a python proteomics framework. *Journal of Proteome Research*, 18(2):709–714, 2019. PMID: 30576148.

[45] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009.

[46] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 07 2008.

[47] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*, 30(10):918–920, 2012.

[48] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[49] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[50] Wout Bittremieux. spectrum_utils: A python package for mass spectrometry data processing and visualization. *Analytical Chemistry*, 92(1):659–661, 2020. PMID: 31809021.

[51] Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, Marc E. Martignoni, Angelika Werner, Rüdiger Hein, Dirk H Busch, Christian Peschel, Roland Rad, Jürgen Cox, Matthias Mann, and Angela M. Krackhardt. Direct

identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. 7:13404.

[52] Alejandro Sanz-Bravo, Carlos Alvarez-Navarro, Adrian Martín-Esteban, Eilon Barnea, Arie Admon, and José A. López de Castro. Ranking the contribution of ankylosing spondylitis-associated endoplasmic reticulum aminopeptidase 1 (ERAP1) polymorphisms to shaping the HLA-b 27 peptidome. 17(7):1308–1323.

[53] Aaron D. Martin, Xueyin Wang, Mark L. Sandberg, Kathleen R. Negri, Ming L. Wu, Dora Toledo Warshaviak, Grant B. Gabrelow, Michele E. McElvain, Bella Lee, Mark E. Daris, Han Xu, and Alexander Kamb. Re-examination of MAGE-a3 as a t-cell therapeutic target. 44(3):95.

[54] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Baozhen Shan, and Ming Li. Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. 2(12):764–771. Number: 12 Publisher: Nature Publishing Group.

[55] Angel Charles, Christopher M. Bourne, Tanya Korontsvit, Zita E. H. Aretz, Sung Soo Mun, Tao Dao, Martin G. Klatt, and David A. Scheinberg. Low-dose CDK4/6 inhibitors induce presentation of pathway specific MHC ligands as potential targets for cancer immunotherapy. 10(1):1916243.

[56] Massimo Andreatta, Annalisa Nicastri, Xu Peng, Gemma Hancock, Lucy Dorrell, Nicola Ternette, and Morten Nielsen. MS-rescue: A computational pipeline to increase the quality and yield of immunopeptidomics experiments. 19(4):1800357. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201800357.

[57] Kiran K. Mangalaparthi, Anil K. Madugundu, Zachary C. Ryan, Kishore Garapati, Jane A. Peterson, Gourav Dey, Amol Prakash, and Akhilesh Pandey. Digging deeper into the immunopeptidome: characterization of post-translationally modified peptides presented by MHC i. 12(3):151–160.

[58] Eric W. Deutsch. File formats commonly used in mass spectrometry proteomics. 11(12):1612–1621.

[59] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022.

[60] Xuan Liu, Congzhi Song, Shichao Liu, Menglu Li, Xionghui Zhou, and Wen Zhang. Multi-way relation-enhanced hypergraph representation learning for anti-cancer drug synergy prediction. 38(20):4782–4789.

[61] Ethan L. Schreiber, Richard E. Korf, and Michael D. Moffitt. Optimal multi-way number partitioning. *J. ACM*, 65(4), jul 2018.

[62] Thomas A. Hansen, Fedor Kryuchkov, and Frank Kjeldsen. Reduction in Database Search Space by Utilization of Amino Acid Composition Information from Electron Transfer Dissociation and Higher-Energy Collisional Dissociation Mass Spectra. *Analytical Chemistry*, 84(15):6638–6645, August 2012. Publisher: American Chemical Society.

[63] R. Stuart Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, 2(3):795–827, 11 2021.

[64] Mingxun Wang, Jian Wang, Jeremy Carver, Benjamin S. Pullman, Seong Won Cha, and Nuno Bandeira. Assembling the community-scale discoverable human proteome. 7(4):412–421.e5. Publisher: Elsevier.

[65] Robert Parker, Arun Tailor, Xu Peng, Annalisa Nicastri, Johannes Zerweck, Ulf Reimer, Holger Wenschuh, Karsten Schnatbaum, and Nicola Ternette. The choice of search engine affects sequencing depth and HLA class i allele-specific peptide repertoires. 20:100124.